International Journal of Computer Science and Engineering (IJCSE)
ISSN 2278-9960
Vol. 2, Issue 4, Sep 2013, 9-16
© IASET

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

A STUDY ON THE IMPACT OF BACKGROUND MODEL IN GMM-UBM BASED SPEAKER VERIFICATION IN MULTI-SENSOR ENVIRONMENT

KSHIROD SARMAH & UTPAL BHATTACHARJEE

Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India

ABSTRACT

In the traditional GMM-UBM based speaker verification (SV) system; it has been observed that the performance of the system depends upon the UBM data selection. Recording environment and specification of the imposter data influences the performance of the system. In this paper, we use two different dataset NIST SRE 2003 and a newly developed database of Arunachali Languages of North East India called Arunachali Language Speech Database (ALS-DB) to analyze the influence of UBM data selection on the GMM-UBM based speaker verification in multi-sensor environment. It has been observed that in multi-sensor environment, performance degradation due to sensor variability is more prominent in comparison to environmental variability of the background model. It has been observed that for same sensor, the average performance degradation due to change in background speaker model is 2.65% in terms of EER whereas for the same background model, the degradation due to sensor variability is 6.34%.

KEYWORDS: GMM-UBM, MFCC, Speaker, Verification

INTRODUCTION

Speaker recognition is a process of determining the identity of a person based on the intrinsic characteristics of his/her utterance or voice. In the source – filter of human speech production, the speech signal is modeled as the convolution output of a vocal source excitation signal and the impulse response of a vocal tract filter system [1]. Linear predictive cepstral coefficients (LPCC) and Mel-frequency cepstral coefficients (MFCC) are the most representative vocal tract –related acoustic and spectral features that modeling the spectral envelope or the formant structure of the vocal tract [2]. Speaker recognition encompasses verification and identification system. Speaker Verification (SV) is the use of a machine to verify the claimed identity of a person from his or her voice. But, in Speaker Identification (SI), the system decides who is the person from his or her utterance [3]. Speaker recognition systems can be divided into text dependent and text-independentones. In text-dependent systems [4] suited for cooperative users, the recognition phrases are fixed, or known beforehand. In text-independent systems, there are no constraints on the words which the speakers are allowed to use [5].

GMM-UBM has become one of the most dominant classification approaches for modeling text-independent speaker recognition application, over the past several decades. In more recent years, GMM-based systems have been applied to the annual NIST Speaker Recognition Evaluations (SRE) [6]. In UBM approach, a single speaker-independent model is trained and then it used for all the speakers in the pool which is one of the most important characteristic of UBM.

Development of automatic, text-independent speaker recognition systems has been seen research into how to get better performance by representing the speaker specific information in speaker models .Probability density functions are used to make representation of the speaker models. For robust speaker modeling, Maximum Likelihood (ML) estimation extends to Maximum A Posteriori (MAP) approach are remarkably used. MAP density estimation techniques incorporate

prior knowledge of how the speaker model parameters vary with respect to the change of recording specification, channel, and recording environments in addition to the speech information provided by the speaker during enrollment [7].

For speaker modeling, the parameters weight, mean and covariance of the UBM are taken as the base model and are adjusted towards the target to speaker model by adapting mean, weight or covariance any one parameter. In lots of literary review of speaker recognition system we observed that the hyper-parameters might be derived from a universal speaker model generally higher Gaussian mixture component order trained from a vast quantity of diverse speech. In this case, [6][13] proposed to utilize a UBM speaker model to derive the relevant adapted speaker models. Thus, the mixture mean prior distribution means are set to the UBM component means.

Many researchers try to select suitable data to improve the quality of UBM in order to get better adapted target speaker models to improve the performance of speaker verification system. In this study, we observe the affect of the performance of speaker verification system due to variation of UBM models trained from different speech databases, ALS-DB and NISTSRE 2003. The rest of the paper is organized as follows. Front-end processing and Overview of GMM-UBM Speaker Modeling are explained in section II and section III. In the section IV gives brief description of Speaker Verification Corpus. Finally Experiments, Result Analysis and Conclusions are explained in section V, section VI and section VII respectively.

FRONT-END PROCESSING

Front-End Processing or feature extraction is also known as speech parameterization. Speech parameterization consists in transforming the speech signal to a set of feature vectors. The purpose of feature extraction phase is to extract the speaker-specific information in the form of feature vectors at reduced data rate which is more compact and more suitable for statistical modeling and the calculation of a distance or any other kind of score. The feature vector represents the speaker–specific information due to vocal tract, excitation source and behavioral traits. A good feature vector set should have representation all of the components of speaker information.

The most representative vocal tract acoustic features are the Linear Predictive Cepstral Coefficients (LPCC) and the Mel Frequency Cepstral Coefficients (MFCC), which aim to extract the speaker vocal tract related and languages related features. Different front-end processing steps applied in the feature extraction method. First, the speech segmented into frames by a 20-ms Hamming window progressing at a 10-ms frame rate. Also, in order to reduce the noise and enhance the high frequency signals, pre-emphasis is adopted with pre-emphasis factor 0.97.A voice activity detector (VAD) is then applied to discard silence-noise frames using energy based criteria.

Next, mel-scale cepstral feature vector are extracted from the speech frames. The mel-scale cepstrum is the discrete cosine transform of the log-spectral energies of the speech segment. The spectral energies are calculated over logarithmically spaced filters with increasing bandwidths (mel-filters). All cepstral coefficients excluding the 0^{th} value which has no speaker specific information as its represents average log-spectral energies are retained in the processing [6]. Finally delta cepstra are computed using first and second order orthogonal polynomial temporal fit over (+2,-2) and (+5,-5) feature vectors from the current vector. Finally, the feature vectors are channel normalized to remove linear channel convolutional effects. Both cepstral mean substraction (CMS) and cepstral variance normalization (CVN) have been used successfully.

In this study, we used only 13 dimensional MFCC features with first and second order derivatives that appended and total of 39 dimensions.

OVERVIEW OF GMM-UBM SPEAKER MODELING

A Universal Background Model (UBM) is one of the most important model used to a biometric verification system to represent general, person-independent feature characteristics to be compared against a model of person-specific feature characteristics when making an accept or reject decision. Most of the modern speaker verification system use a UBM for modeling the alternative hypothesis in the likelihood ratio test [6].

The GMM-UBM can be seen as a likelihood-ratio detector, where the UBM is trained to represent the speaker-independent distribution of features and the GMM is adapted from the UBM using MAP algorithm to ideal speaker model containing individual speaker characteristics. In this GMM-UBM system, a UBM is firstly trained to capture the general gender independent characteristic of all the speakers (other than the target speakers). The UBM parameters include weights, mean vectors and covariance matrices, which can be expressed as

$$\lambda=\{w_m,\mu_m,\sum_m\}_{m=1}^{M} \quad (1)$$

Where, M is the number of Gaussian mixtures. In speaker recognition, usually the value of M is large (1024 in this case) and the covariance matrices are often set in diagonal form, which makes the fast computation.

The theory explains for determining the statistic from a single feature vector observation sample from the target or non-target speaker classes. This test statistic deals with two speaker classes identified as the target speaker and non-target (UBM) speaker set specified by models, λ_{targ} and λ_{ub}.

For a given T independent and identically distributed observations, X= $\{x_1, x_2, x_3, \ldots\ldots x_T\}$. The joint likelihood ratio may be determined. A more robust of measure for speaker verification is the expected frame-based log-likelihood ratio measure can be defined as follows.

$$E[LLR(X)] = E[logP(X|\lambda_{\text{target}}) - logP(X|\lambda_{\text{ubm}})] \quad (2)$$

$$E[LLR(X)] = \frac{1}{T}\sum_{t=1}^{T}(\log p(x_t|\lambda_{\text{target}}) - \log p(x_t|\lambda_{\text{ubm}})) \quad (3)$$

The UBM is a large GMM trained to represent the speaker-independent distribution of features. To train a UBM, the simplest approach is to merely pool all the data and use it to train the UBM via the EM algorithm. We should be careful that the pooled data is balance of male and female speech to create a gender independent UBM model [6].

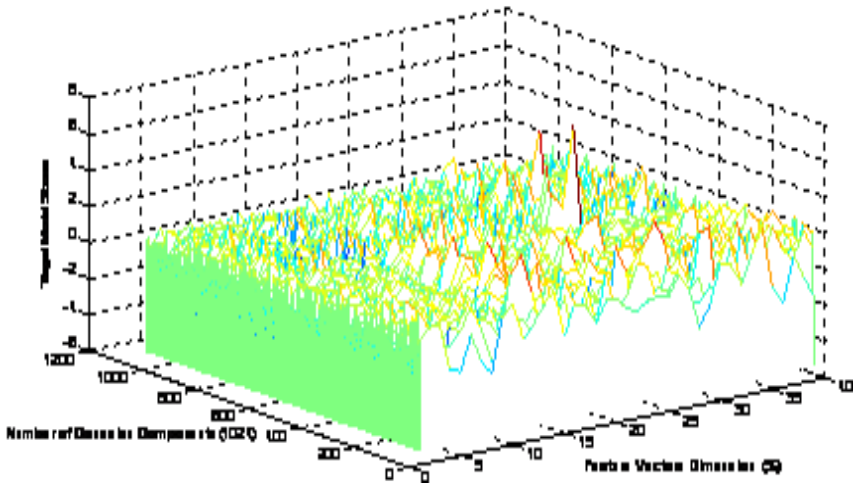

Figure 1: The MAP (Mean) Adapted Target Speaker Model Scores for a Speaker Training from ALS-DB and UBM Model from the Same ALS-DB Database

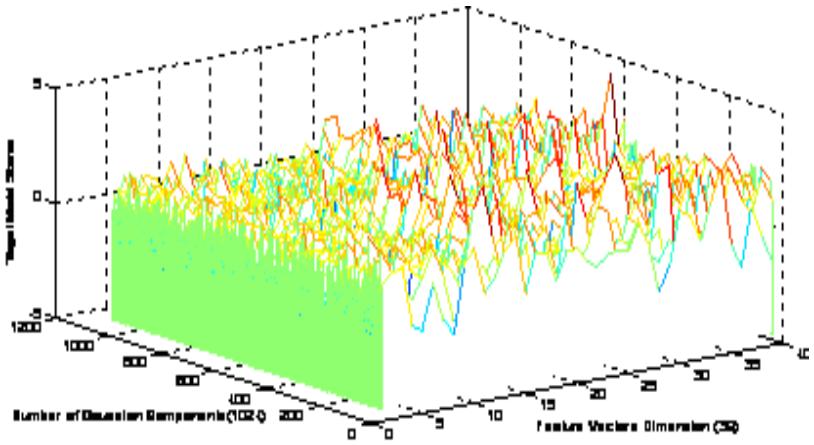

Figure 2: The MAP (Mean) Adapted the Same Target Speaker Model Scores for the Same Speaker Training from ALS-DB and UBM Model from NISTSRE 2003 Database

From the above first diagram (Figure 1), it is clear that the score of Gaussian components of the target speaker model that adapted (mean variable only) from the UBM of the database ALS-DB which has smooth variation due matching condition of speaker modeling data with the UBM data. But in the second diagram (Figure 2) it is showed that the score is highly varied for the same speaker model that adapted from the UBM thatconstructed from NISTSRE2003 for the mismatching data of speaker model and UBM modeling.

SPEAKER VERIFICATION CORPUS

In this section we used the recently collected Arunachali Language Speech Database (ALS-DB) [11] [12] and NISTSRE2003 standard database.

To study the impact of language variability and channel variability on speaker recognition task, ALS-DB is collected in multilingual environment. Each speaker is recorded for three different languages – English, Hindi and a Local language, which belongs to any one of the four major Arunachali languages - Adi, Nyishi, Galo and Apatani. Each recording is of 4-5 minutes duration. Speech data were recorded in parallel across four recording devices, which are listed in table -1.

Table 1: Device and Recording Specifications

Device Sl. No	Device Type	Sampling Rate	File Format
Device 1 (D1)	Table mounted microphone	16 kHz	wav
Device 2 (D2)	Headset microphone	16 kHz	wav
Device 3 (D3)	Laptop microphone	16 kHz	wav
Device 4 (D4)	Portable Voice Recorder	44.1 kHz	mp3

The speakers are recorded for reading style of conversation. The speech data collection was done in laboratory environment with air conditioner, server and other equipments switched on. The speech data was contributed by 100 male and 80 female informants chosen from the age group 20-50 years. During recording, the subject was asked to read a story from the school book of duration 4-5 minutes in each language for twice and the second reading was considered for recording. Each informant participates in four recording sessions and there is a gap of at least one week between two sessions.

For UBM modeling 100 male and 100 female non-speakers are collected from the database with same enrollment duration. The experiments are carried out for all four devices D1, D2, D3 and D4 separately for training the speaker models as well as creating the UBM with the same experimental set-up.

For the other UBM with the same number 1024 Gaussian components constructed from the database NISTSRE 2003. In this case also we collect the equal number of speech samples from 200 speakers (100 male and 100 female), which were automatically imposters for the target speaker models trained from the database ALS-DB.

EXPERIMENTS

In this works, the baseline system of the speaker verification system was developed using Gaussian Mixture Model with Universal Background model (GMM-UBM) based modeling approach. A 39-dimensional feature vector was used, made up of 13 mel-frequency cepstral coefficient (MFCC) and their first order derivatives as well as second order derivatives. The first order derivatives were approximated over three samples and for second over five samples.

The coefficients were extracted from a speech sampled at 16 KHz with 16 bits/sample resolution. A pre-emphasis filter $H(z)=1- 0.97z^{-1}$ has been applied before framing. The pre-emphasized speech signal is segmented into frame of 20 ms with frame rate 10ms. Each frame is multiplied by a Hamming window. From the windowed frame, FFT has been computed and the magnitude spectrum is filtered with a bank of 22 triangular filters spaced on Mel-scale and constrained into a frequency band of 300-3400 Hz. The log-compressed filter outputs are converted to cepstral coefficients by DCT.

The 0^{th}cepstral coefficient is not used in the cepstral feature vector since it corresponds to the energy of the whole frame and only 12 MFCC coefficients have been used [10]. To capture the time varying nature of the speech signal, the first order and second order derivative of the Cepstral coefficients are also calculated. Combining the MFCC coefficients with its first order and second derivatives, so finally we get a 36-dimensional feature vector. Cepstral mean subtraction has been applied on all features to reduce the effect of channel mismatch.

The Gaussian mixture model with 1024 Gaussian components has been used for both the UBM and speaker model. The UBM was created by training the speaker model with 50 male and 50 female speaker's data with 512 Gaussian components each male and female model with Expectation Maximization (EM) algorithm. Finally gender-independent UBM model is created by pooling the both male and female models of total 1024 Gaussian components. The speaker models were created by adapting only the mean parameters of the UBM using maximum a posteriori (MAP) approach with the speaker specific data.

The detection error trade-off (DET) curve has been plotted using log likelihood ratio between the claimed model and the UBM and the equal error rate (EER) obtained from the DET curve has been used as a measure for the performance of the speaker verification system.

Another measurement Minimum DCF values has also been evaluated.

The performance measures are the same as NIST speaker recognition evaluation (NIST, 2010), using the equal error rate (EER) and minimum detection cost function (DCF).

According to the NIST Detection Cost Function (DCF) can be defined as

$$DCF = 0.1P_{Miss} + 0.99\ P_{FA} \quad (4)$$

Where $P_{Miss}$and $P_{FA}$are the miss (false rejection) probability and the false alarm (false acceptance) probability respectively.

Minimum DCF (MinDCF), defined as the DCF value at the threshold for which DCF value is smallest, is the optimum cost.

RESULT ANALYSIS

Case 1: In this experiment, UBM is trained using the imposter's utterances from the database ALS-DB for all Devices D1, D2, D3 and D4. That means in this case both target speaker model and UBM model were constructed from the same database ALS-DB.

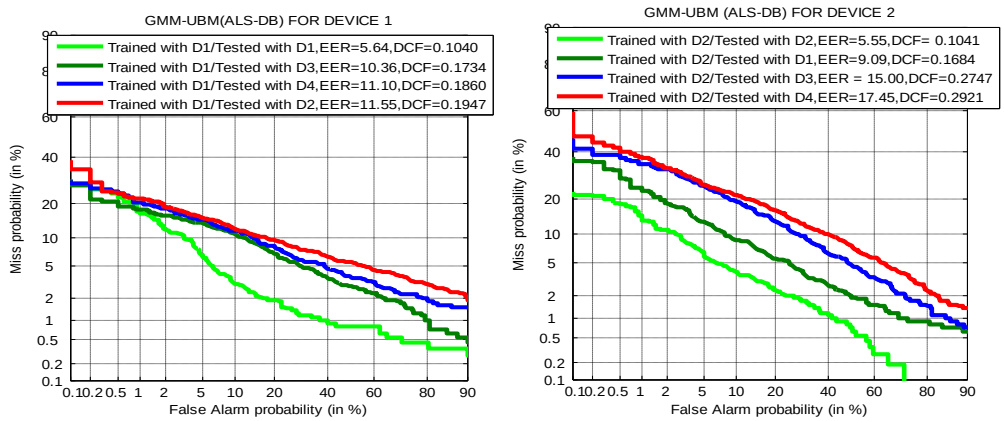

Figure 3(a),(b): The DET Curves for the GMM-UBM Speaker Verification System Using the UBM Model from the Database ALS-DB Trained with D1and D2 and Testing with D1, D2, D3, and D4 Devices Separately

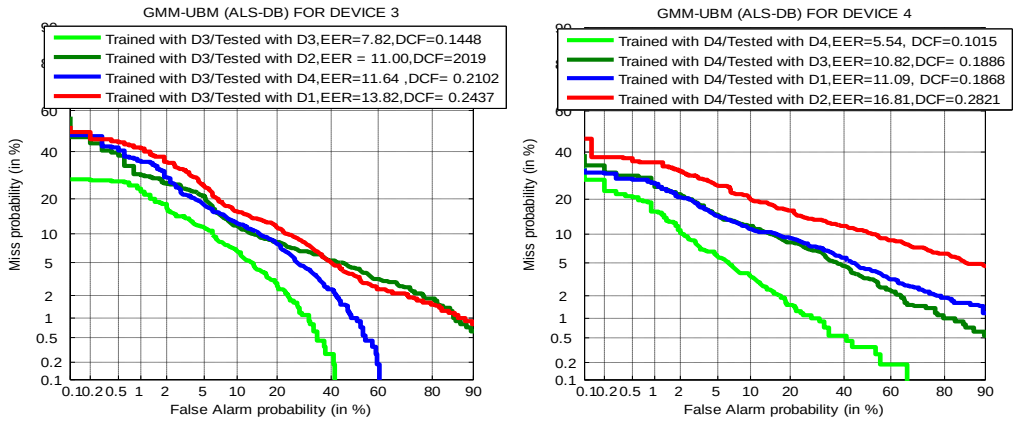

Figure 3(c),(d): The DET Curves for the GMM-UBM Speaker Verification System Using the UBM Model from the Database ALS-DB Trained with D3 and D4 Testing with D1, D2, D3, and D4 Devices Separately

Case 2: In this experiment, the target speaker modeling is constructed from the database ALS-DB for all Devices D1, D2, D3 and D4 but the UBM is trained using the imposter's utterances from the database NISTSRE2003.

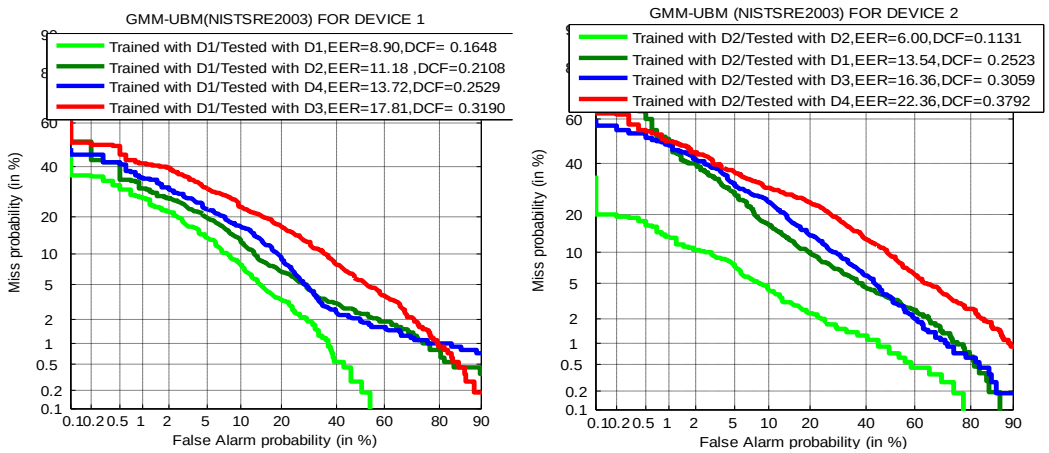

Figure 4(a),(b): The DET Curves for the GMM-UBM Speaker Verification System Using the UBM Model from the Database NISTSRE2003 Trained with D1 and D2 Testing with D1, D2, D3, and D4 Devices Separately

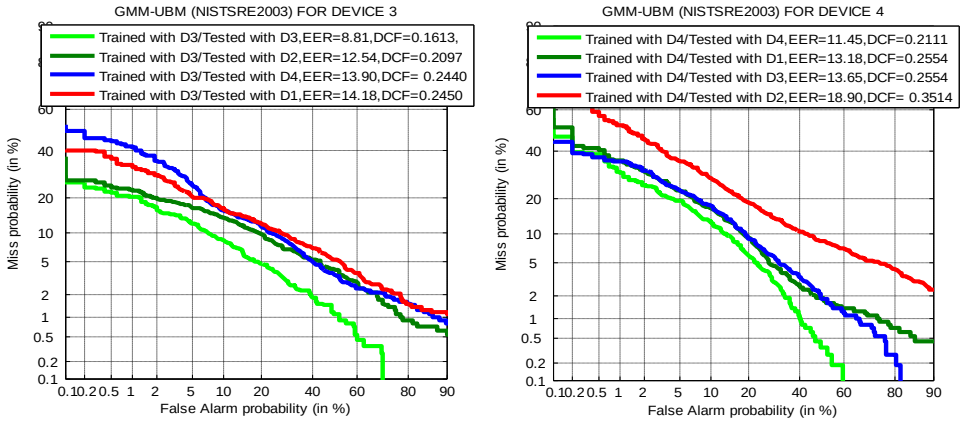

Figure 4(c),(d): The DET Curves for the GMM-UBM Speaker Verification System Using the UBM Model from the Database NISTSRE2003 Trained with D3 and D4 Testing with D1, D2, D3, and D4 Devices Separately

Table 2: The EER and Min DCF Values of Speaker Verification System

Training Devices	Testing Devices	GMM-UBM(ALS-DB)		GMM-UBM(NISTSRE2003)	
		EER%	Min DCF	EER%	Min DCF
D1	D1	5.64	0.1040	8.90	0.1648
	D2	11.55	0.1947	11.18	0.2108
	D3	10.36	0.1734	17.81	0.3190
	D4	11.10	0.1860	13.72	0.2529
D2	D1	9.09	0.1684	13.54	0.2523
	D2	5.55	0.1041	6.00	0.1131
	D3	15.00	0.2747	16.36	0.3059
	D4	17.45	0.2921	22.36	0.3792
D3	D1	13.82	0.2437	14.18	0.2450
	D2	11.00	0.2019	12.54	0.2097
	D3	7.82	0.1448	8.81	0.1613
	D4	11.62	0.2102	13.90	0.2440
D4	D1	11.09	0.1868	13.18	0.2554
	D2	16.81	0.2821	18.90	0.3514
	D3	10.82	0.1868	13.65	0.2554
	D4	5.54	0.1015	11.45	0.2111

CONCLUSIONS

From the above experiment point of view we have observed that the performance of the speaker verification system degrades for mismatching condition of training and testing data for different sensors. The performance of SV system with respect to the UBM created from NIST SRE 2003 degrades of approximately **3.58%, 1.0%, 1.5% and 6.00%** of EER values that of the ALS-DB database for the Device D1, D2, D3 and D4 respectively. So, we conclude that the performance of GMM-UBM based SV system highly dependent not only the speaker modeling data but also on the qualities of imposter's data and its recording environments that used for building UBM models. The performance degrades highly for highly mismatched data between speaker modeling and UBM modeling. Finally, it has been observed that for same sensor, the average performance degradation due to change in background speaker model is **2.65%** in terms of EER whereas for the same background model, the degradation due to sensor variability is **6.34%.** From the above experiment, it becomes clear that the degradation is more prominent in case of sensor variability then background speaker model variability.

ACKNOWLEDGEMENTS

This work has been supported by the ongoing project grant No. 12(12)/2009-ESD sponsored by the Department of Information Technology, Government of India.

REFERENCES

1. L.R. Rabiner&R.W.Schafer Digital Processing of Speech Signals, Englewood Cliffs, NJ: Prentice-Hall. 1978
2. S.Furui, "Cepstral analysis technique for automatic speaker verification", *IEEE Trans. Acoust., Speech, Signal Process*., vol. ASSP-29,no.2,pp 254-272. April 1981.
3. J.P.Campbell, "Speaker Recognition: a tutorials" *Proc. IEEE*, vol. 85, no.9, pp. 1437-1462, Sep 1997..
4. M. H´ebert, J. Benesty M. Sondhi and Y.Huang, "Text-dependent speaker recognition". In *Springer handbook of speech processing* SpringerVerlag, pp. 743–762, 2008.
5. Tomi. Kinnunen, " An Overview of Text-Independent Speaker Recognition: from Features to Supervectors", *Speech Communication* 52(1):12-40,2010.
6. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "SpeakerVerification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol. 10(1–3), pp. 19-41,2000.
7. Jason. Pelecanos, Robbie. Vogt, &Sridha. Sridharan, "A study on standard and iterative MAP adaptation for speaker recognition", *Proceeding on the 9th Australian International Conference on Speech Science & Technology* Melbourne, Dec. 2002.
8. D.A. Reynolds," Experimental evaluation of features for robust speaker identification", *IEEE Trans. Speech Audio Process*., vol 2(4), pp. 639-643,Oct. 1994.
9. D.A. Reynolds, Universal Background Models*", MIT Lincoln Laboratory,244 wood St. Lexinton, MA 02140,USA.
10. Z. Xiaojia, S. Yang & W. DeLiang, "Robust speaker identification using a CASA front-end", *Acoustics, Speech and Signal Processing (ICASSP),* IEEE International Conference on, pp.5468-5471,2011.
11. UtpalBhattacharjee, &KshirodSarmah, "A Multilingual Speech Database for Speaker Recognition", *Proc. IEEE, ISPCC.* June 2012.
12. UtpalBhattacharjee, &KshirodSarmah, "Development of a Speech Corpus for Speaker Verification Research in Multilingual Environment", *International Journal of Soft Computing and Engineering (IJSCE)* ISSN: 2231-2307, Volume-2, Issue-6,pp 443-446, January 2013.
13. D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification", In *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech '97)*, Rhodes, Greece, vol. 2, pp. 963–966,Sep. 1997.